X_400004 - Statistics
Final Resit

9 February 2021

Instructions:

- The exam is to be solved **individually**.
- Please **write clearly and in an organised way**: we can't grade illegible answers.
- Please change pages when starting a new question.
- This is an exam on a mathematical subject, so support your answers with **computations**, rather than words, whenever possible.
- You should report **all relevant computations** and **justify** non-trivial steps.
- This is a **closed book exam**; you are only allowed to have one A4 sheet with **handwritten** notes with you.
- You may use a calculator.
- There are 7 pages in the exam questionnaire (including this one) and you have two hours (120 minutes) to complete the exam.
- Students entitled to extra time have an extra 20 minutes.
- The exam consists of 11 questions spread throughout 3 problems.
- The number of points per question is indicated next to it for a total of 100 points.
- Your final grade is $\max(1, \text{score}/10)$, where "score" is the number you points you get.
- The problems are not necessarily ordered in term of difficulty. I recommend that you quickly read through all problems first, then do the problems in whatever order suits you best.
- Remember to **identify** at least the first page of your answer sheets with you name, course name, and student number.

After completing your exam, digitalise your answer sheets (with the correct order and orientation) and submit them as **a single PDF** on Canvas for grading. **You have 10 minutes following the end of the exam to do this** after which any submission will be marked as late. These instructions do not replace the VU's *Protocol of online examination for 2020-2021* that you can find on Canvas.

Prob.I: Suppose that you are managing an email mailing list for a website. Let $X_1, \ldots, X_n$ be a random sample from the Bernoulli distribution with unknown parameter $p \in [0, 1]$. Each observation X_i corresponds to a different person that got the mailing list email. Suppose that $X_i = 1$ if the person opened the email, and $X_i = 0$, otherwise. You are interested in the value of p, the probability that someone who gets the mailing list actually opens it. (The other parameter in the model, n, is not specified but you always know what it is.)

8 pts (a) Describe a near-pivot T for p. Justify how you arrived at T, and don't forget to mention the (approximate) distribution of the pivot.

12 pts (b) Use the near-pivot from the previous question to derive a confidence interval of level 0.95 for p. Don't just present the final result: show how you go from (near-)pivot to confidence interval.

7 pts (c) Suppose now that the last issue of the mailing list was sent to $n = 27$ subscribers and that, out of these, only 2 opened the email. Based on the confidence interval for p that you found, what would be your most optimistic guess as to how large p is? (There are tables at the end of the exam questionnaire in Appendix B that you may need to consult to answer this question.)

10 pts (d) Suppose that you are not happy with the large amount of uncertainty in the confidence interval that you got – you find the confidence interval too wide. In class, we saw that increasing the sample size n typically reduces uncertainty. How many emails would you have to send to be fairly confident that the estimate $\bar{X}$ of p is not off by more that 0.005? (You can still use $\bar{X}$ as your best guess for p.)

Prob.II: When studying the browsing behaviour in websites you come across the following statistical problem: users accessing a webpage will take some amount of time until they click the button "continue". Somewhat simplistically, the amount of time (in seconds) a user takes to press the button is well modeled by an exponential distribution with parameter λ (and therefore expectation $1/\lambda$). Suppose you have observations from n users, denoted by $X_1, \ldots, X_n$, that can be assumed to be an i.i.d. sample from an exponential distribution with parameter λ.

We would like to conduct the following hypothesis test

$$H_0 : \lambda = 0.25 \quad \text{against} \quad H_1 : \lambda < 0.25 \ .$$

In other words, is the average time a user stays on the page $1/0.25 = 4$ seconds, or is it larger? A natural test statistic to consider is $Y = \sum_{i=1}^{n} X_i$. From your knowledge of probability you know that Y has an Erlang distribution with parameters n and λ.

Hint: An Erlang random variable Y with parameters n and λ has density

$$f_Y(y) = \begin{cases} e^{-\lambda y} \frac{\lambda^n y^{n-1}}{(n-1)!} & \text{if } y \geq 0 \\ 0 & \text{otherwise} \end{cases} ,$$

and cumulative distribution function

$$F_Y(y) = P(Y \leq y) = \begin{cases} 1 - e^{-\lambda y} \sum_{k=0}^{n-1} \frac{(\lambda y)^k}{k!} & \text{if } y \geq 0 \\ 0 & \text{otherwise} \end{cases} .$$

Consider a test procedure that rejects the null hypothesis if $Y \geq c_\alpha$, where $c_\alpha > 0$ must be chosen depending on the desired significance level. **For the rest of the question consider the case $n = 3$.**

11 pts (a) Say that you take $c_\alpha = 32$. What is the type I error of this test?

11 pts (b) For a poorly designed webpage we expect the users to spend on average 16 seconds, meaning $\lambda = 1/16$. What is the power of the test in the previous question (where $c_\alpha = 32$) in this case?

11 pts (c) An experiment was conducted and a total waiting time of $y = 28$ seconds was recorded. Compute the p-value of this test. Would you reject the hull hypothesis at significance level $\alpha = 0.05$? Carefully justify your answer.

Prob.III: Consider the following situation. You are spending some well deserved holidays on the beach, in a country with consistently good weather. While relaxing on the sand you notice that the number of swimmers in the water seems heavily influenced by the water temperature. This prompts the question: can you use the number of swimmers as a "thermometer"?

Over the course of two weeks you count the number of swimmers entering the water between 10:30 and 11:00 in a pre-selected region of the beach. In addition, you take note of the water temperature (as reported by a local meteorological site). Suppose that the data below is what you collected.

Day	Number of swimmers	Water temperature (°C)
1	57	19.0
2	88	19.5
3	89	23.5
4	73	17.5
5	91	21.0
6	100	20.0
7	70	21.5
8	109	22.0
9	101	25.0
10	91	20.5
11	79	22.0
12	96	23.5
13	82	22.5
14	101	23.5

A simple linear regression analysis is to be conducted where the water temperature is taken as the response variable and the number of swimmers as the predictor. **Have a look at Appendix A before you start solving this problem.**

4 pts (a) Write the assumed **model equation** for the relation between the number of swimmers and the water temperature using α to denote the intercept and β the slope. Suppose that you fit the model and get the plot in Figure 2 for the resulting residuals. Is it reasonable to assume normally distributed errors?

11 pts (b) Estimate the model parameters α and β from the data, as well as σ^2, the variance of the noise.

4 pts (c) Suppose that after a quick trip to the country side you returned to the same beach and decided to see if your model was indeed useful in predicting the water temperature. Between 10:30 and 11:00 you counted 93 swimmers. Give an estimate for the water temperature (according to your model).

11 pts (d) Your friends were impressed by your model, but a bit doubtful about the quality of your estimate in (c) and wanted to have a better idea of the errors that are involved. Use your knowledge of regression models to give a two sided **prediction interval** for the water temperature in the scenario in (c) (use $\alpha = 0.05$).

A Beach Data

From the dataset we get $\bar{x} = 87.64286$, $\bar{y} = 21.5$, $\sum_{i=1}^{n} x_i^2 = 110169$, $\sum_{i=1}^{n} y_i^2 = 6527$, $\sum_{i=1}^{n} x_i y_i = 26585$.

Below follows a normal QQ-plot of the residuals of the model.

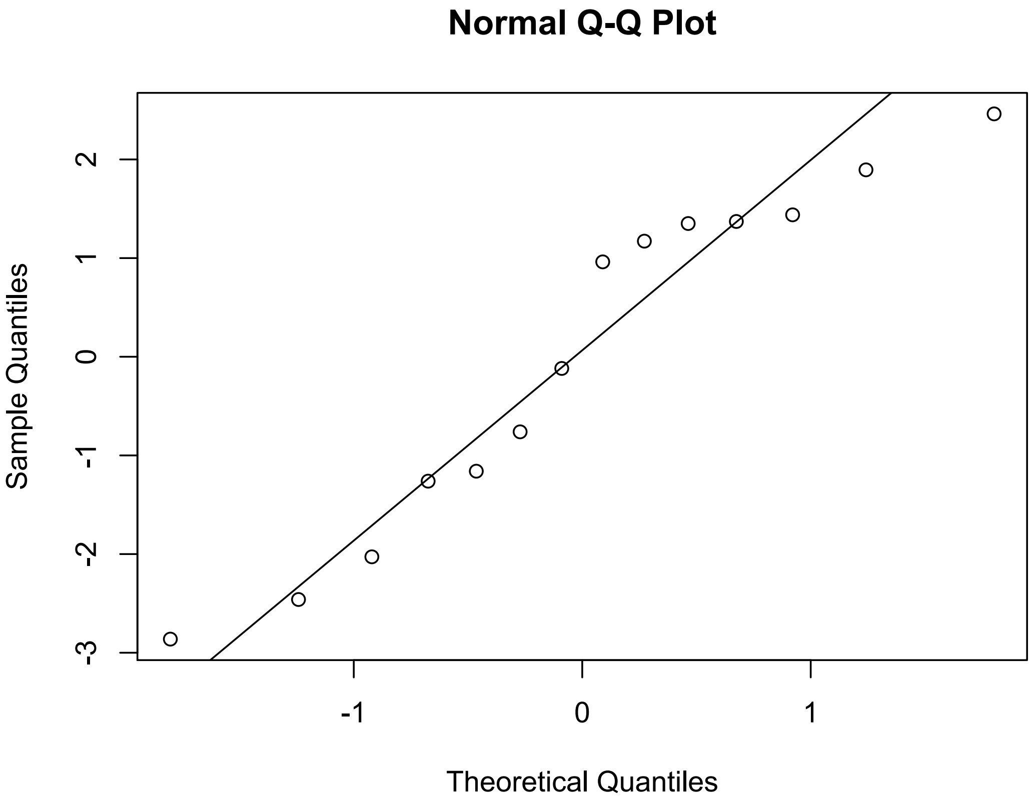

Figure 1: Normal QQ plot of the residuals of the regression model.

B Quantiles and cumulative probability tables

Below you can find a table of probabilities for the t-distribution and for the standard normal distribution, respectively. Please read the caption of the tables carefully.

$\downarrow \nu/\alpha \rightarrow$	0.3	0.2	0.15	0.1	0.05	0.025	0.02	0.01	0.005	0.0025	0.001
1	0.727	1.376	1.963	3.078	6.314	12.71	15.90	31.82	63.66	127.3	318.3
2	0.617	1.061	1.386	1.886	2.920	4.303	4.849	6.965	9.925	14.10	22.33
3	0.584	0.978	1.250	1.638	2.353	3.182	3.482	4.541	5.841	7.453	10.215
4	0.569	0.941	1.190	1.533	2.132	**2.776**	2.999	3.747	4.604	5.598	7.173
5	0.559	0.920	1.156	1.476	2.015	2.571	2.757	3.365	4.032	4.773	5.893
6	0.553	0.906	1.134	1.440	1.943	2.447	2.612	3.143	3.707	4.317	5.208
7	0.549	0.896	1.119	1.415	1.895	2.365	2.517	2.998	3.499	4.029	4.785
8	0.546	0.889	1.108	1.397	1.860	2.306	2.449	2.896	3.355	3.833	4.501
9	0.543	0.883	1.100	1.383	1.833	2.262	2.398	2.821	3.250	3.690	4.297
10	0.542	0.879	1.093	1.372	1.812	2.228	2.359	2.764	3.169	3.581	4.144
11	0.540	0.876	1.088	1.363	1.796	2.201	2.328	2.718	3.106	3.497	4.025
12	0.539	0.873	1.083	1.356	1.782	2.179	2.303	2.681	3.055	3.428	3.930
13	0.538	0.870	1.079	1.350	1.771	2.160	2.282	2.650	3.012	3.372	3.852
14	0.537	0.868	1.076	1.345	1.761	2.145	2.264	2.624	2.977	3.326	3.787
15	0.536	0.866	1.074	1.341	1.753	2.131	2.249	2.602	2.947	3.286	3.733
16	0.535	0.865	1.071	1.337	1.746	2.120	2.235	2.583	2.921	3.252	3.686
17	0.534	0.863	1.069	1.333	1.740	2.110	2.224	2.567	2.898	3.222	3.646
18	0.534	0.862	1.067	1.330	1.734	2.101	2.214	2.552	2.878	3.197	3.610
19	0.533	0.861	1.066	1.328	1.729	2.093	2.205	2.539	2.861	3.174	3.579
20	0.533	0.860	1.064	1.325	1.725	2.086	2.197	2.528	2.845	3.153	3.552
21	0.532	0.859	1.063	1.323	1.721	2.080	2.189	2.518	2.831	3.135	3.527
22	0.532	0.858	1.061	1.321	1.717	2.074	2.183	2.508	2.819	3.119	3.505
23	0.532	0.858	1.060	1.319	1.714	2.069	2.177	2.500	2.807	3.104	3.485
24	0.531	0.857	1.059	1.318	1.711	2.064	2.172	2.492	2.797	3.091	3.467
25	0.531	0.856	1.058	1.316	1.708	2.060	2.167	2.485	2.787	3.078	3.450
26	0.531	0.856	1.058	1.315	1.706	2.056	2.162	2.479	2.779	3.067	3.435
27	0.531	0.855	1.057	1.314	1.703	2.052	2.158	2.473	2.771	3.057	3.421
28	0.530	0.855	1.056	1.313	1.701	2.048	2.154	2.467	2.763	3.047	3.408
29	0.530	0.854	1.055	1.311	1.699	2.045	2.150	2.462	2.756	3.038	3.396
∞	0.524	0.842	1.036	1.282	1.645	1.960	2.054	2.326	2.576	2.807	3.090

Figure 2: Right quantiles of the student t-distribution with ν degrees of freedom. Example: if X is a student t distributed random variable, with $\nu = 4$ degrees of freedom then $P(X \geq 2.776) = 0.025$. **The entries of the table are therefore $t_{\nu,1-\alpha}$.**

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7703	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	**0.9115**	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3.0	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990
3.1	0.9990	0.9991	0.9991	0.9991	0.9992	0.9992	0.9992	0.9992	0.9993	0.9993
3.2	0.9993	0.9993	0.9994	0.9994	0.9994	0.9994	0.9994	0.9995	0.9995	0.9995
3.3	0.9995	0.9995	0.9995	0.9996	0.9996	0.9996	0.9996	0.9996	0.9996	0.9997
3.4	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9998
3.5	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998
3.6	0.9998	0.9998	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999
3.7	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999
3.8	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999
3.9	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000

Figure 3: Cumulative probabilities of the standard normal distribution. The rows correspond to the number rounded down to the closest decimal, and columns represent the second decimal. Example: if Z is a standard normal random variable then $P(Z \leq 1.35) = 0.9115$. You look this number up in the row corresponding to 1.3, and in the 6th column (corresponding to 0.05.) Note that for $z < 0$ we have $P(Z \leq z) = 1 - P(Z \leq -z)$. So for example $P(Z \leq -1.35) = 1 - 0.9115 = 0.0885$.